Application of Text Analysis Techniques for Alignment of Ontological Categories

Daichi Shirayama and Keiichi Nakata

***Abstract*—Ontologies provide powerful means for structuring and accessing information. To increase the effectiveness in using multiple ontologies and to cater for the proliferation of ontologies, alignment and traversal of multiple ontologies have gained importance. We report on the application of similarity matching based on text analysis techniques to identify corresponding nodes between multiple ontology trees, in this case categories in search engines. By this method, we address not only the problem of matching concepts but also identifying concepts in different ontologies that may share the same label but have different semantic scope.**

***Index Terms*—Categorization, information systems, ontology engineering, text processing.**

I. INTRODUCTION

In today's "information society", we are flooded with numerous categories of information, which is often presented in a mixed form. In order to find answers for a particular information need, we should be able to select necessary information efficiently. In recent years, ontologies are often used to express knowledge required for problem solving and organize information systematically. With the help of ontologies, we can classify knowledge and information systematically and intelligibly, thereby increasing the accessibility of information sources.

However, the construction of ontologies are often time-consuming and labor intensive, and even for the same domain, they tend to differ according to definitions employed for concepts represented, i.e., they cannot be uniquely constructed. Moreover, contents of an information source may change over time and an ontology would ideally reflect such changes.

To deal with these problems, there could be three types of approaches:

- Development of collaborative ontology editors. [1,2]
- Automatic ontology construction.
- Extension of existing ontologies. [3]

All of these avenues are being pursued in the domain of ontology engineering. In this paper, we describe an approach in the third type. We believe that a system that automatically identifies the similarities and differences between ontologies would be instrumental in enabling reuse of existing ontologies. Hence our objective is the development of a system that semi-automatically finds alignment between two ontologies. While there are many different interpretations and examples of ontologies, we assume that the type of ontologies we deal with here will have the following characteristics:

1. An ontology describes taxonomic relations between concepts.
2. Each concept in an ontology is explicated by a set of texts.

The first assumption supposes that an ontology is a hierarchical description of concepts. This differs from the knowledge-based view employed in AI which expects a frame-like structure. While taxonomic ontologies are less descriptive, it has been used in many applications, including information retrieval. The category-style Web portals and search engines such as Yahoo, library classification systems, and XML tags defined by DTDs and RDFs, can be seen as examples of this type of ontology. Both types of ontologies describe how a particular domain can be conceptualized, reflecting the way how the domain is captured, or "understood", by the developers/users of the ontology.

Upon such observations, the second assumption above becomes reasonable. Concepts in such ontologies are not usually *defined*, i.e., there is no explicit definition of each concept in the ontology. For instance, a library category of "International Politics" assumes that which books are to be classified into this category is obvious to the librarian. This means that the concept of "International Politics" is "defined" by the contents of books that are classified into this category. The same applies for search engine categories.

The basic operations that may be performed to ontologies in this context are *merging* and *alignment*. Here we subscribe to Noy et al. [3] for their definitions. Merging refers to the operation of uniting two or more ontologies. When two ontologies are aligned, the corresponding nodes (in various levels of similarities) are identified (Fig.1). Hence, alignment can be seen as a precondition for merging ontologies.

There are many "clues" for finding alignments, e.g., node labels, keywords and data structures. However, in our work we intentionally disregard node labels, since, while they provide obvious correspondence between nodes, they can be sometimes misleading. For example, the same linguistic label may be referring to different concepts (homonyms). Also even if they

The authors are affiliated to the Institute of Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan (phone/fax: +81-3-5841-8087, email: sirayama@cse.k.
u-tokyo.ac.jp, nakata@k.u-tokyo.ac.jp).

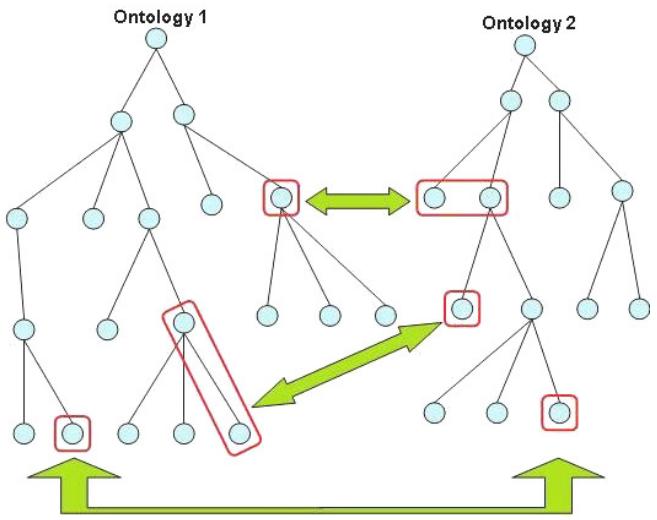

Fig.1. Alignment of two ontologies. Aligments may be one-to-one, many-to-one, or, in a more complex cases, many-to-many.

refer to the same concept, their scope might be different. These are problems that occur since concept labels are often chosen arbitrarily and subjectively. Instead, we analyze the set of texts that represent a concept and calculate similarities between nodes from two ontologies based on document set similarities. A similar approach of comparing two ontological classes is taken in [4], but we concentrate more on finding alignments based on this information.

II. Method

A. Document characterization

In order to characterize a concept described by a set of texts, we apply text analysis techniques [5] to extract the elements (words) that express the contents of a document. Words that describe the contents of a document are generally referred to as *index words*. That is, the contents of the document are approximated by the set of the index words extracted out of the document. For European languages, words that consists a document apart from "stop words" are candidates for index words. In Japanese, since words are not separated by spaces, the morphological analysis is necessary to identify word boundaries. Moreover, nouns are often chosen for index words, since they are more descriptive than other parts of speech such as verbs and adjectives.

B. Text feature vectors

Coverage of a document by a set of index words can be characterized by calculating the weight of words based on how often a word appears in the text. The term frequency (TF) of an index word w_t that appears in document d is defined as follows:

$$w_t^d = \frac{F_i}{F}$$

Here, F is the total number of words in document d, and F_i is the number of occurrence of the word w_t in document d. When an individual document is characterized, we used a list of meaningless words that consists of words that are too general and therefore not helpful and words that are identified as such by the morphology analyzer but do not exist, and removed them from the list of index words.

Based on term frequencies, a vector whose element position identifies an index word and its value being its term frequency can be generated. This can be regarded as a feature vector that describes the document.

To compare the similarity between two concepts, we calculate the similarity of the feature vectors that are created from the set of texts (documents) that describe each concept. The c*osine similarity* is often used as a measure of similarity between two feature vectors.

C. Finding corresponding nodes

To align two ontologies O_1 and O_2, we attempt to identify one or more nodes in O_2 that correspond a node in O_1. We treat this as a search problem that finds the most similar node in an ontology (the search tree). Since we can evaluate the similarity between the nodes as described in the previous subsection, we can apply an informed search method for this purpose.

III. Experiment

A. Problem specification

To evaluate the applicability of this approach, we attempted to perform the alignment of two commercial search engines that provide directory-style categories, *viz.*, Yahoo Japan (http://www.yahoo.co.jp) and Lycos Japan (http://www.lycos.co.jp). These were chosen, since, 1) the Web page categorization can be viewed as a taxonomic ontology, 2) each category (concept node) contains Web pages that characterize it, and 3) these ontologies cover the same domain, i.e., the domain of general Web pages, and they do not differ so much but have enough differences that require human judgment as to whether two nodes should be aligned.

The procedure employed by the system is as follows:

1) A node (category) in Lycos is selected (*input node*) for which one of more corresponding nodes in Yahoo are to be identified.
2) The feature vector of a node is constructed by collecting the Web pages in that category. Here, pages are collected through the link analysis of pages that describe the node.
3) The top page of each URL that appears in 2) are collated and treated as one document. The morphological analysis is performed using *Chasen* [6], followed by the removal of meaningless words to create a list of index words.
4) The TF value of each element in the list of index words is calculated, and the feature vector (TF vector) of the node is created.
5) A corresponding node in Yahoo is determined by performing hill-climbing search through the Yahoo directory tree starting from the root ("Home"), using the cosine similarity between the input node and the Yahoo node as the evaluation function.

In order to perform step 2 efficiently, a robot (web crawler),

which crawls the Web automatically by following URL links was also developed. The system was implemented in Perl.

B. Preliminary experiment

To check whether cosine similarity measures of nodes' feature vectors are meaningful, we first conducted a preliminary experiment. For this purpose, the cosine similarity between the TF vector of a category node in Infoseek Japan (http://www.infoseek.co.jp)

home > health & beauty > diet fitness > diet

and 50 random nodes of Yahoo were compared.

The result is shown in Fig.2. This shows that three categories at the very left have significantly high similarity values compared to other nodes. This means that we can differentiate between similar and non-similar nodes effectively using this measure.

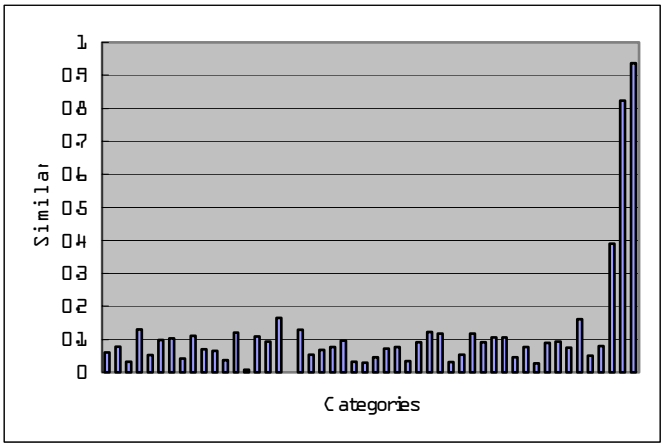

Fig.2. Similarity values for a Infoseek category "diet" and a set of randomly selected Yahoo categories.

C. Experiment and results

We used the system developed to automatically search the Yahoo node that corresponds to an input node selected from Lycos.

For each element in the test set of 21 input nodes in Lycos, we first asked human subjects to identify "correct" nodes and a set of "partially correct" nodes. Then the system performed search to identify the corresponding node for each input node, and the result was evaluated by comparing it to the "correct" and "partially correct" nodes. The result is shown in Table 1.

TABLE 1.
MATCHING RESULT OF A LYCOS (JAPAN) CATEGORY TO YAHOO (JAPAN) CATEGORIES (CATEGORY ONLY)

Lycos category (input)	Search result of the Yahoo category	Simil-arity	Result
Lycos/directory/medical treatment, health/sickness, and symptom/cancer	Home > entertainment > SF, fantasy, and horror	0.6104	
Lycos/directory/lifestyle/fortune-telling/style water	Home > Business and Economy > shopping and service > entertainment > fortune-telling > style water	0.8897	C
Lycos/directory/hobby and sports/hobby/gambling/totocalcio	Home > news and media > television > program > sports	0.3893	
Lycos/directory/society and mass communication/sociology/women's studies	Home > life and culture > sex > gender free	0.5025	C
Lycos/directory/society and mass communication/publication/magazine/opinion magazine	Home > news and media > magazine	0.3331	PC
Lycos/directory/education/school/special school/design and dress	Home > education > vocational schools > fashion, dress and housekeeping	0.8351	C
Lycos/directory/hobby and sports/baseball/National High School Baseball Championship	Home > entertainment > entertainer > celebrity	0.1973	
Lycos/directory/regional information/hot spring information/secret hot spring	Home > recreation and sports > travel and sightseeing > tour guide > hot spring	0.7149	C
Lycos/directory/natural science, technology/resource, and energy/nuclear power	Home > natural science and technology > energy > nuclear power	0.8038	C
Lycos/directory/computer Internet/Internet/Homepage authoring/HTML	Home > Business and Economy > traffic > timetable	0.1429	
Lycos/directory/economy and investment and industry/market	Home > Business and Economy > finance and investment > news and media	0.767	C
Lycos/directory/medical treatment, health/nursing, and nursing	Home > regional information > links	0.4421	
Lycos/directory/politics, law/law, and jurisprudential/constitution	Home > politics > method > constitution	0.9805	C
Lycos/directory/natural science and technology/environment/environmental protection	Home > natural science and technology > ecology	0.4475	C
Lycos/directory/economy and industry/consumer	Home > entertainment	0.4493	
Lycos/directory/natural science and technology/living thing science/genetics	Home > regional information	0.4441	
Lycos/directory/society, mass communication/disaster, and disaster prevention	Home > Business and Economy > wanted.	0.3811	
Lycos/directory/natural science and technology/paranormal phenomena	Home > social science > periodical > journals	0.6564	
Lycos/directory/natural science and technology/mathematics	Home > social science > sociology > sociology of education	0.1982	
Lycos/directory/arts and literature science/language study and linguistics/English	Home > news and media > international news	0.309	
Lycos/directory/lifestyle/interpersonal relationship/meeting/lover wanted.	Home > Business and Economy > shopping and service > entertainment > meeting	0.7839	C

TABLE 2.
MATCHING RESULT OF A LYCOS (JAPAN) CATEGORY TO YAHOO (JAPAN) CATEGORIES (CONSIDERING HIGHER CATEGORY INFORMATION)

Lycos category (input)	Search result the Yahoo category	Simil-arity	Result
Lycos/directory/medical treatment, health/sickness, and symptom/cancer	Home > health and medical science > dentistry	0.6235	
Lycos/directory/lifestyle/fortune-telling/style water	Home > entertainment > advice > fortune-telling > directories	0.2976	PC
Lycos/directory/hobby and sports/hobby/gambling/totocalcio	Home > news and media > television > program > sports	0.3893	
Lycos/directory/society and mass communication/sociology/women's studies	Home > life and culture > sex > gender free	0.5025	C
Lycos/directory/society and mass communication/publication/magazine/opinion magazine	Home > Business and Economy > wanted.	0.3516	
Lycos/directory/education/school/special school/design and dress	Home > education > vocational schools >fashion, dress and housekeeping	0.8351	C
Lycos/directory/hobby and sports/baseball/National High School Baseball Championship	Home > recreation and sports > sports > news and media	0.3531	
Lycos/directory/regional information/hot spring information/secret hot spring	Home > recreation and sports > travel and sightseeing > tour guide > hot spring	0.7149	C
Lycos/directory/natural science, technology/resource, and energy/nuclear power	Home > natural science and technology > energy > directories	0.6284	PC
Lycos/directory/computer Internet/Internet/Homepage authoring/HTML	Home > computers and the Internet > standard	0.3043	
Lycos/directory/economy and investment and industry/market	Home > Business and Economy > finance and investment > news and media	0.767	C
Lycos/directory/medical treatment, health/nursing, and nursing	Home > regional information > links	0.4421	
Lycos/directory/politics, law/law, and jurisprudential/constitution	Home > politics > method > constitution	0.9805	C
Lycos/directory/natural science and technology/environment/environmental protection	Home > natural science and technology > ecology	0.4475	C
Lycos/directory/economy and industry/consumer	Home > entertainment	0.4493	
Lycos/directory/natural science and technology/living thing science/genetics	Home > natural science and technology > research	0.5302	
Lycos/directory/society, mass communication/disaster, and disaster prevention	Home > Business and Economy > wanted.	0.3811	
Lycos/directory/natural science and technology/paranormal phenomena	Home > natural science and technology > group	0.5988	
Lycos/directory/natural science and technology/mathematics	Home > natural science and technology > mathematics > mathematician	0.8458	C
Lycos/directory/arts and literature science/language study and linguistics/English	Home > social science > linguistics > language > English > group > ESS	0.3975	PC
Lycos/directory/lifestyle/interpersonal relationship/meeting/lover wanted.	Home > Business and Economy > shopping and service > entertainment > meeting	0.7839	C

Here, "C" indicates that the node found by the system was a "correct" node, and "PC" indicates that the node found was one of the nodes in the "partially correct" nodes. The result shows that in 9 out of 21 cases the system could identify the correct node and in 1 case a partially correct node. It also shows that while most correct nodes have high similarity values, this is not always the case.

IV. DISCUSSION

The result shown in Table 1 indicates that the precision of this method was not as high as it could be desired, and in more than half of the cases it returned results considered to be "wrong" by human subjects. It also seems that a better result for this test set could have been obtained if a straightforward node label matching, with a help of a thesaurus, was used. However, in some cases it identified a correspondence between nodes which can only be found upon inspection of their contents. For instance, a simple label matching would not find an alignment between a Lycos category:

society and mass communication/sociology/women's studies

and a Yahoo category:

life and culture > sex > gender free

This is an example of term mismatch, which may be solved by an elaborate thesaurus. On the other hand, another example of correspondence between a Lycos category:

natural science and technology/environment/environmental protection

and a Yahoo category:

Home > natural science and technology > ecology

involves the difference in the scope of the concept, or the granularity of conceptual units, employed in these particular ontologies. Such a mismatch between concept labels and actual data cannot be correctly identified by node label matching only, even with a help of a thesaurus.

Upon inspection of the first result, we hypothesized that we could obtain better results by taking into consideration the information stored in parent nodes. The algorithm was slightly modified to combine the TF vector of a node with that of its parent nodes with a reduction coefficient of 0.5 for each upper (ISA) link followed. This emulates the activation propagation in semantic nets. The result is shown in Table 2.

In comparison to the first result, the number of "correct" cases decreased by one with 8 nodes out of 21. However, the number of cases evaluated as "partially correct" increased to three, giving an impression that the performance had improved overall. The number of cases tested was not enough to conclude whether the results obtained were significant.

We observe that the reasons that this method worked well for some examples are as follows:

- The categories in the search engines we examined were suitably represented by the documents that are classified into them.
- The two ontologies (categorizations) were similar enough so that a suitable alignment could be found by applying search starting from the root node.
- The two ontologies were reasonably structured so that it represented the conceptual hierarchies appropriately, albeit differences in categorizations and conceptual scoping.

On the other hand, the we observed that the reasons for wrongly identified alignment are as follows:

- The search method was a simple hill-climbing search which did not allow backtracking, thereby leading some searches to local maxima.
- The cosine similarity of TF vectors used to guide search was in some cases inappropriate.
- Some categories had Web pages that contained little text but mainly pictures and scripts, which may have affected the contents of the TF vector.

To overcome the first problem above, we can consider using a search algorithm that reduces the chance of falling into local maxima such as best-first search, or that allows backtracking. However, the cost of retrieving Web pages and constructing TF vectors dynamically is high and such search algorithms may face the performance problem. Since our problem domain was search engine categorizations which are very large and are subject to dynamic changes caused by the addition of new pages in the Internet, preprocessing of node data was not suitable. In other domains which are smaller in size and more static, a more elaborate search method should offer higher precision.

V. Conclusion

We reported an approach of using text analysis techniques to the ontology alignment problem, and demonstrated its feasibility by applying it to Web page categorizations of two commercial Web search engines. This method intentionally ignored the usage of category labels, since the choice of such labels are often arbitrary. The result obtained by using a simple hill-climbing search using the cosine similarity of term frequency vectors that represent category nodes suggests that while a better precision was desirable, it found alignments between categories which would not have been found by an alignment strategy that relies primarily on label matching.

Further research, including experimentations on more cases and different ontologies, is necessary to make any concrete conclusions concerning the general applicability of this method. A systematic way to combine this method and other alignment techniques should be explored, in order to develop effective automatic ontology alignment tools.

Acknowledgment

This work is supported by CREST Program, Japan Science and Technology Corporation (JST).

References

[1] A. Farquhar, R. Fikes, W. Pratt, J. Rice, “Collaborative Ontology Construction for Information Integration,” Knowledge Systems Laboratory, Department of Computer Science, Stanford University, KSL-95-63, 1995.

[2] S. Decker, M. Erdmann, D. Fensel, and R. Studer, “Ontobroker: Ontology based access to distributed and semi-structured information,” in R. Meersman et al., editor, DS-8: Semantic Issues in Multimedia Systems, Kluwer Academic Publisher, 1999 pp. 351-369.

[3] N. Fridman Noy, M. A. Musen:, “PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment,” in AAAI/IAAI 2000, 2000, pp.450-455.

[4] M. S. Lacher, G. Groh, “Facilitating the exchange of explicit knowledge through ontology mappings,” 14th International and FLAIR Conference, AAAI Press, 2001.

[5] G. Salton. “Automatic Text Processing,.” Addison-Wesley Publishing, Reading, Mass., 1989.

[6] Nara Institute of Science and Technology, 1997 http://chasen.aist-nara.ac.jp/